Identification of "Missing Link" Families of Small DNA Tumor Viruses

Nicole L. Welch[1], Michael J. Tisza[1], Anna Belford[1], Diana V. Pastrana[1], Yuk-Ying S. Pang[1], John T. Schiller[1], Ping An[2], Paul G. Cantalupo[2], James M. Pipas[2], Samantha Koda[3], Kuttichantran Subramaniam[3], Thomas B. Waltzek[3], Chao Bian[4], Qiong Shi[4], Zhiqiang Ruan[4], Terry Fei Fan Ng[5], Gabriel J. Starrett[1], and Christopher B. Buck[1]*

[1]Lab of Cellular Oncology, NCI, NIH, Bethesda, MD, 20892

[2]Dept. of Biological Sciences, Univ. of Pittsburgh, Pittsburgh, PA 15260

[3]Dept. of Infectious Disease and Pathology, Univ. of Florida, Gainesville, FL, 32611

[4]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI, Shenzhen, Guangdong 518083, China

[5]Division of Viral Diseases, NCIRD, CDC, Atlanta, GA 30333

*Corresponding author: buckc@mail.nih.gov

Abstract

Adenoviruses, papillomaviruses, parvoviruses and polyomaviruses are collectively known as small DNA tumor viruses. Although it has long been recognized that small DNA tumor virus oncoproteins and capsid proteins show a variety of structural and functional similarities, it is unclear whether these similarities reflect descent from a common ancestor, convergent evolution, horizontal gene transfer among virus lineages, or repeated acquisition of genes from host cells. We report the discovery of fourteen new members of an emerging fourth family of small DNA tumor viruses, the *Adomaviridae*. Cell culture-based expression of adomavirus open reading frames showed that the virion structural proteins of adomaviruses are homologous to those of adenoviruses. The search for adomaviruses unexpectedly revealed the existence of a previously unrecognized family of animal DNA viruses that are distantly similar to virophages and other polinton-like viruses of unicellular eukaryotes. Members of the new "adintovirus" family encode retrovirus-like integrase proteins and adenovirus-like DNA polymerase and virion structural proteins. Adintovirus sequences were found in genomics datasets for all major groups of animals, including terrestrial vertebrates. Analysis of the gene content of adintoviruses supports a model in which small DNA tumor viruses descended from more complex adintovirus-like ancestors through gene-loss and horizontal gene transfer events occurring hundreds of millions of years ago.

Author Summary

In contrast to cellular organisms, viruses do not encode any universally conserved genes. Even within a given family of viruses, the primary amino acid sequences of homologous genes can rapidly diverge to the point of unrecognizability. Although modern metagenomic surveys have begun to expose the wide sequence diversity within known virus families, ascertaining whether the large amount of unidentifiable sequence "dark matter" typically observed in these surveys represents unknown virus groups remains a substantial challenge. Our searches of sequence databases revealed more than a dozen highly divergent new members of an emerging virus family, the *Adomaviridae*. Using methods for detection of distant homologs, we also discovered the existence of an apparently older lineage of viruses that we name "adintoviruses." Examples of adintoviruses, which combine a retrovirus-like integrase gene with adenovirus-like DNA polymerase and virion structural proteins, can be found as integrants or as apparently linear DNA molecules in genome sequencing datasets for all major animal phyla. The results suggest a framework that ties small DNA tumor viruses into a shared evolutionary history in which key family-level divergence events occurred in early animals and their unicellular precursors.

Introduction

Polyomaviruses, papillomaviruses, parvoviruses, and adenoviruses are collectively referred to as small DNA tumor viruses. The four virus families encode non-enveloped virion proteins with similar β-jellyroll core folds and express functionally similar oncogenes that inactivate cellular tumor suppressor proteins (Figure 1)(*1, 2*). It has remained unclear how small DNA tumor viruses might be interrelated. A current evolutionary model has placed adenoviruses on a spectrum of double-stranded DNA (dsDNA) polinton-like viruses that encode hallmark type B DNA polymerase (PolB) and integrase genes. Polyomaviruses, papillomaviruses, and parvoviruses are proposed to have descended from unrelated circular Rep-encoding single-stranded DNA (CRESS) virus ancestors (*3*). While this model accurately represents many of the features of modern small DNA tumor viruses, it does not explain a number of observed similarities between the four families. Achieving a better understanding of the evolutionary origins of small DNA tumor viruses has the potential to direct comparative studies of these ubiquitous human pathogens.

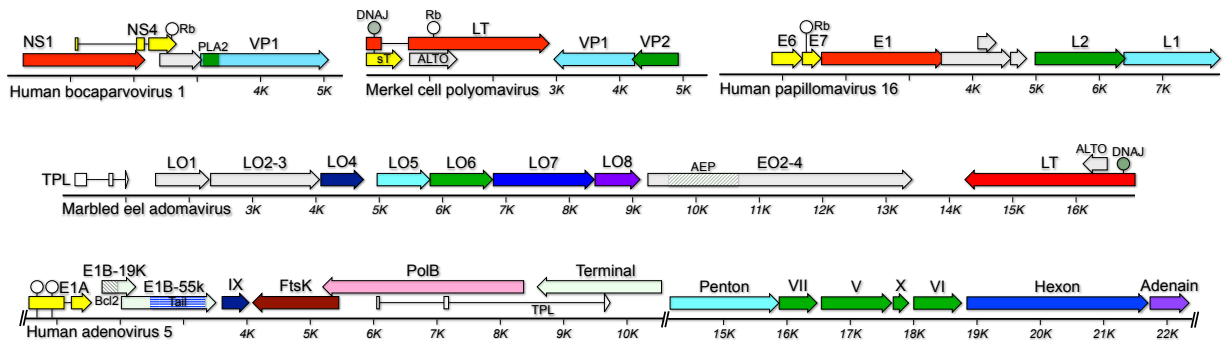

Figure 1: Examples of small DNA tumor virus genomes. Genes are color-coded based on known or inferred functions. AEP: domain resembling archaeal-eukaryotic primase small catalytic subunit; white lollipop: retinoblastoma (Rb) interaction motif; gray lollipop: domain with predicted fold similar to DNAJ chaperone proteins; PLA2: domain with predicted fold similar to phospholipase A2; Tail: domain with predicted structural similarity to dsDNA virus tail proteins or carbohydrate degrading enzymes. See main text for explanation of other gene names. Supplemental Table 1 lists accession numbers and host animal taxonomic designations.

In 2011, a novel circular dsDNA virus was discovered in a lethal disease outbreak among Japanese eels (*Anguilla japonica*)(*4, 5*). Two related viruses have since been isolated from Taiwanese marbled eels (*Anguilla marmorata*) and giant guitarfish (*Rhynchobatus djiddensis*)(*6, 7*). The name "adomaviruses" has been applied to this emerging family based on the fact that they encode proteins with distant similarity to the virion-maturational proteases of *ad*enoviruses as well as proteins distantly similar to the replicative superfamily 3 DNA helicases (S3H) of poly*oma*viruses, papill*oma*viruses, and parvoviruses (*7*).

The starting goal of the present study was to identify the adomavirus major and minor capsid proteins. To discover additional adomaviruses, particularly those that might serve as laboratory models for adomavirus infection, we performed metagenomic sequencing and developed a bioinformatics pipeline to detect candidate adomavirus sequences in the NCBI's Sequence Read

Archive (SRA). In addition to identifying more than a dozen previously unknown adomaviruses, the searches serendipitously revealed the existence of a previously unrecognized virus family that encodes distinctive PolB and virion proteins that are distantly similar to those of adenoviruses. Members of the previously unrecognized family also encode a retrovirus-like integrase gene that is distantly similar to the integrase genes of a group of unicellular eukaryote-associated polinton-like viruses known as virophages (*8*). Analysis of the gene content of the emerging "adintovirus" family sheds light on the evolutionary history of small DNA tumor viruses.

Results

Identification of adomavirus virion proteins

In a gene naming convention developed for Japanese eel adomavirus (NC_015123), hypothetical proteins are numbered using inferred late-strand open reading frame (LO) and early-strand open reading frame (EO) designations. HHPred searches revealed that adomavirus LO8 proteins resemble the virion maturational proteases of adenoviruses. This suggests the hypothesis that adomavirus LO4-7 genes might be syntenic homologs of adenovirus virion structural proteins. We set out to experimentally test this hypothesis.

To determine whether predicted adomavirus LO proteins are expressed from spliced or unspliced ORFs, RNAseq data published by Wen and colleagues (*6*) were analyzed to determine the splicing patterns of marbled eel adomavirus transcripts. Splice acceptors immediately upstream of the inferred ATG initiator codons of LO4, LO5, LO6, and LO7 proteins were extensively utilized (Supplemental Table 2). Messenger RNAs encoding the adenovirus late genes carry a tripartite leader (TPL) that has been shown to enhance translation (*9*). A similar three-exon leader sequence was detected upstream of the marbled eel adomavirus LO genes (Figure 1).

HHpred searches were performed to detect predicted structural similarities between adomavirus and adenovirus proteins. LO6 ORFs show high-probability (>95%) predicted structural similarity to a 37 amino acid hydrophobic segment of adenovirus pX, a minor virion core protein that is thought to participate in condensation of the viral chromatin (*10*). LO6 alignments also showed moderate probability hits for adenovirus pVI, which is believed to play a role in destabilizing cellular membranes during the infectious entry process (*11*). The results suggest that LO6 is a fused homolog of adenovirus pVI and pX virion core proteins.

Adomavirus LO4 proteins have a predicted C-terminal trimeric coiled-coil domain. This relatively generic predicted fold gives a large and diverse range of hits in HHpred searches, including the coiled-coil domains of various viral fiber proteins (for example, avian reovirus σC). Negative-stain EM findings indicate that adomavirus virions do not have a discernible vertex fiber (*4, 6, 7*). Some individual LO4 proteins, as well as alignments of LO4 proteins, yielded HHpred hits for adenovirus pIX, a trimeric coiled coil protein that serves as a "cement" that smooths the triangular facets of the adenovirus virion. The hypothesis that LO4 is a pIX homolog could explain the smooth appearance of adomavirus facets in negative-stain EM.

Although LO5 and LO7 did not yield informative results in BLAST, HHpred, or Phyre[2] searches, the inferred homology of the adomavirus LO4, LO6, and LO8 genes to minor components of adenovirus virions led us to hypothesize that LO5 is a syntenic homolog of

122 adenovirus penton (a pentameric single-jellyroll protein that occupies virion vertices) and that
123 LO7 is a syntenic homolog of adenovirus hexon (a tandem double-jellyroll protein that
124 trimerizes to form virion facets) (*1, 10*). To experimentally test this prediction, marbled eel
125 adomavirus was grown in eel kidney cell culture (*6*) and virions were purified using Optiprep
126 gradient ultracentrifugation. Virion-rich gradient fractions were separated on SDS-PAGE gels
127 and bands were subjected to mass spectrometric analysis (Supplemental Figure 1). The results
128 indicate the presence of substantial amounts of actin and myosin derived from host eel cells.
129 Adomavirus LO2-3 proteins show high probability HHpred hits for Wiscott-Aldrich syndrome
130 protein 1 and other actin-regulatory proteins, suggesting the hypothesis that the virion
131 preparations might be contaminated with host cell actin due to interactions with LO2-3. Only a
132 small number of individual peptides representing LO2-3 were detected by mass spectrometry.
133 The analysis identified prominent bands in the Coomassie-stained gel as LO4, LO5, LO7, and
134 LO8. The relative intensities of bands identified as LO5 and LO7 are consistent with the
135 prediction that the two proteins constitute penton and hexon subunits, respectively. Lower
136 molecular weight bands showed hits for LO6, consistent with the prediction that this protein is
137 present in virions in an LO8-cleaved form. Marbled eel and guitarfish adomavirus LO6 proteins
138 encode potential adenain cleavage motifs ((MIL)XGXG or L(LR)GG)(*12*). A list of protein
139 modifications observed in the mass spectrometric analysis is shown in Supplemental Table 3.

140 Adenovirus penton proteins can spontaneously assemble into 12-pentamer subviral particles that
141 may serve as decoy pseudocapsids in vivo (*13*). To investigate the possibility of a comparable
142 phenomenon in adomaviruses, codon-modified marbled eel adomavirus LO1-LO8 expression
143 plasmids were transfected individually into mammalian 293TT cells (*14*). Optiprep
144 ultracentrifugation was used to separate virus-like particles from smaller solutes. A human
145 papillomavirus type 16 (HPV16) L1/L2 expression plasmid was used as a positive control for
146 virus-like particle (VLP) formation (*15*). Cells transfected with adomavirus LO4, LO5, or LO7
147 expression constructs each showed robust production of particles that migrated into the core
148 fractions of Optiprep gradients (Figure 2), whereas cells transfected with LO1, LO2-3, LO6, or
149 LO8 alone did not show evidence of particle formation (Supplemental Figure 2). In co-
150 transfections of various combinations of LO genes, it was found that inclusion of LO6 inhibited
151 the formation of LO5 and LO7 particles but did not impair the formation of LO4 particles
152 (Supplemental Figure 3). This result suggests that LO6 is a minor virion component that directly
153 interacts with LO5 and LO7.

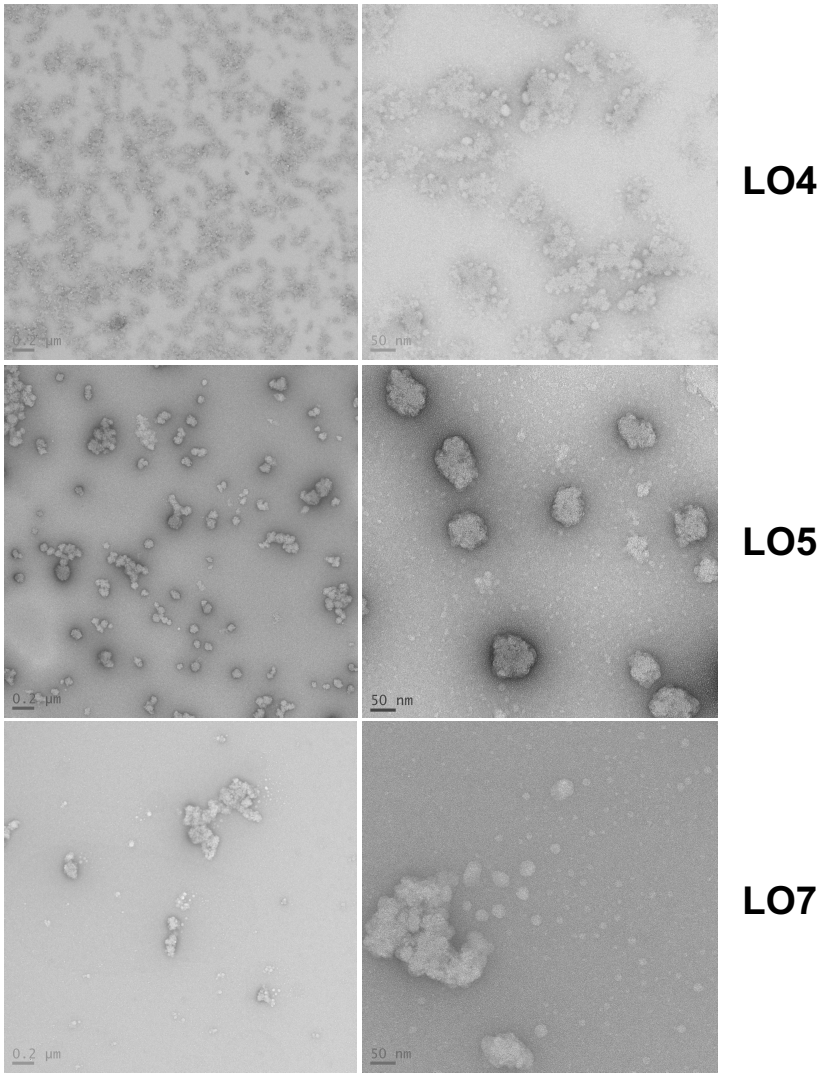

Figure 2. Negative stain EM of particles formed after expression of single LO proteins in mammalian cells. Mammalian 293TT cells were transfected with expression plasmids encoding the indicated marbled eel adomavirus proteins. Particles were extracted from cells by detergent lysis and purified by Optiprep ultracentrifugation. Scale bars are shown at the bottom left corner of each electron micrograph.

Individually expressed LO5 and LO7 particle preparations showed dsDNA signal in Quant-iT PicoGreen assays (Invitrogen), indicating the presence of nuclease-resistant encapsidated DNA within the purified particles. Optiprep-purified particle preparations from cells co-transfected with LO4, LO5, and LO7 were subjected to an additional round of nuclease digestion with salt-tolerant Benzonase endonuclease (Sigma) followed by agarose gel filtration to remove the nuclease and any residual digested DNA fragments. The gel filtered particles reproducibly contained roughly seven nanograms of DNA per microgram of total protein, confirming the presence of nuclease-resistant nonspecific cellular DNA within the particles.

163 Taken together, these results support the hypothesis that LO5, LO6, LO7, and LO8 are syntenic
164 homologs of adenovirus penton, pX/pVI, hexon, and adenain, respectively. LO4 also appears to
165 be a virion structural protein but the question of whether it serves as a facet cement will require
166 further investigation.
167

168 **Detection of additional adomaviruses**

169 In an effort to discover adomaviruses that infect animals amenable to laboratory study, we
170 performed sequencing of aquarium fish specimens and pursued data mining of public sequence
171 repositories. Post-mortem metagenomics analysis of an aquarium-bred Amazon red discus
172 cichlid (*Symphysodon discus*) exhibiting lethargy and inflamed skin lesions revealed a circular
173 ~20 kb adomavirus genome. Histopathological analysis of skin lesions from the discus specimen
174 showed no evidence of intranuclear inclusions or other obvious histopathology.

175 Another adomavirus from an apparently healthy green arowana (*Scleropages formosus*) was first
176 identified as a set of short NCBI Whole-Genome Shotgun (WGS) contigs with similarity to
177 Japanese eel adomavirus. A complete adomavirus genome was characterized by Sanger
178 sequencing of overlapping PCR products using DNA left over from the original fin snip used for
179 the WGS project (*16*).

180 A bioinformatics pipeline was developed to probe the SRA. Datasets that appeared to be rich in
181 adomavirus-like sequences based on translated nucleotides were subjected to de novo assembly.
182 This approach resulted in the identification of 10 complete or nearly complete adomavirus
183 genomes representing distinct species (Supplemental Table 1, Figure 3). Notably, two
184 adomaviruses were identified from genome sequencing datasets for farmed tilapia (*Oreochromis*
185 *niloticus*), which represent a $9 billion per year global industry. Adomavirus-like fragments were
186 also detected in a brook salamander (*Calotriton asper*) and two closely related species of anole
187 lizard (genus *Anolis*). Our pipeline did not have adequate throughput to comprehensively search
188 the much larger volume of sequence data for mammals and birds.

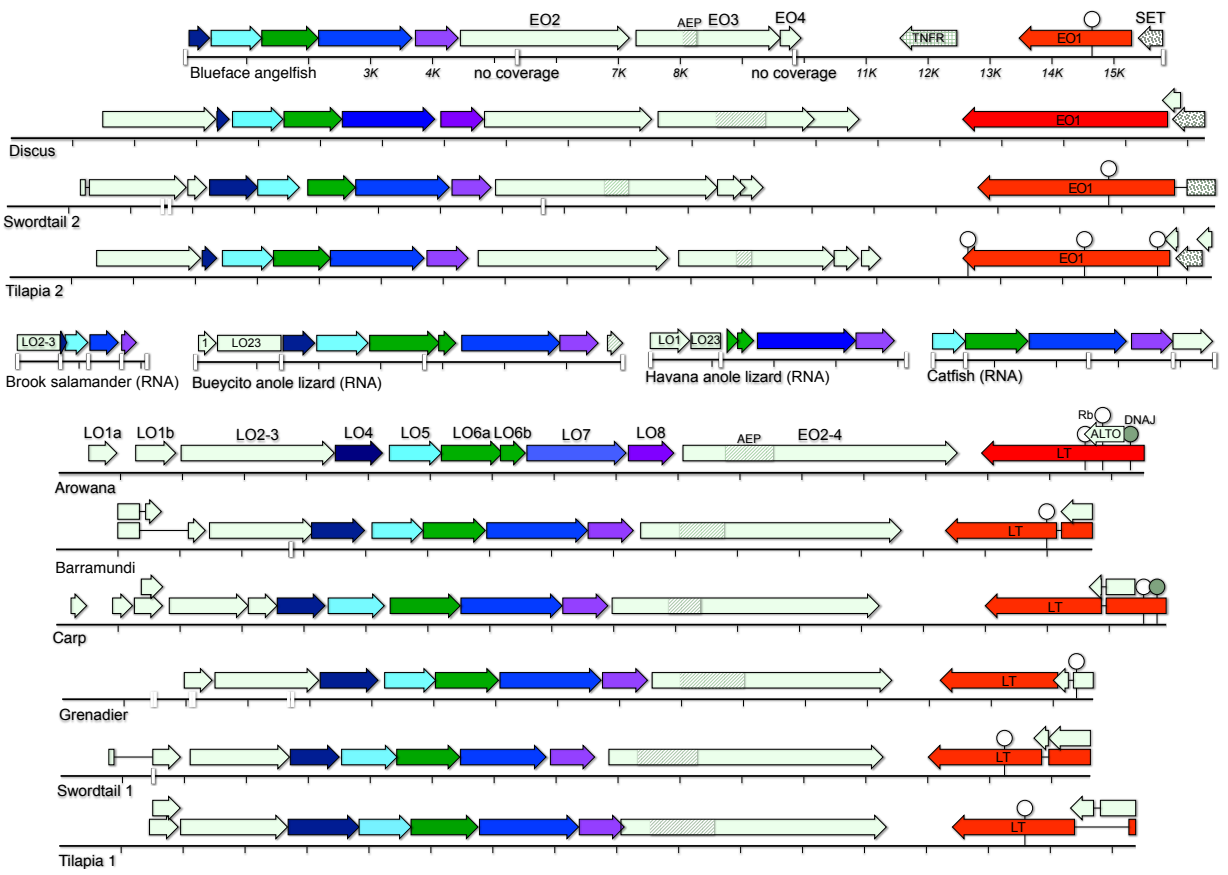

Figure 3: Picture maps of previously unknown adomaviruses. The figure uses the same color scheme as Figure 1. Alpha adomaviruses are defined by the presence of an adoma-specific superfamily 3 helicase (S3H) gene, designated EO1. Beta adomaviruses instead encode a polyomavirus large T antigen (LT). In many cases, repetitive or GC-rich patches (particularly near the 3' end of the LO2-3 ORF) could not be resolved using available short-read datasets. Coverage gaps are represented as white bars. A poorly conserved set of accessory genes upstream of beta adomavirus EO1 genes show varying degrees of similarity to the S-adenosyl methionine-binding pocket of cellular SET proteins, which function as histone lysine methyltransferases. Adomavirus SET homologs are highly divergent from all previously described eukaryotic and viral SET genes. The same is true for adomavirus EO2-4 and EO3 segments that encode homologs of the catalytic small subunit of archaeal eukaryotic primases (AEPs). Supplemental Table 1 lists virus accession numbers and Linnaean designations of host animals.

Adomavirus sequences can be divided into two groups based on their replicative S3H genes. A group we designate alpha is defined by the presence of an EO1 replicase gene distantly similar to papillomavirus E1 proteins (Figure 4). Sequence and predicted structural similarities include a C-terminal S3H domain and a central nicking endonuclease-like segment referred to as the origin binding domain (*3, 17, 18*). Members of adomavirus clade beta encode proteins highly similar to the large T antigen (LT) proteins of polyomaviruses. The adomavirus LT-like proteins share a full range of familiar polyomavirus-like features, including an N-terminal DnaJ domain, a potential retinoblastoma-interaction motif (LXCXE or LXXLFD) (*19*), a nickase-like origin-binding domain, and a C-terminal S3H domain (*20*).

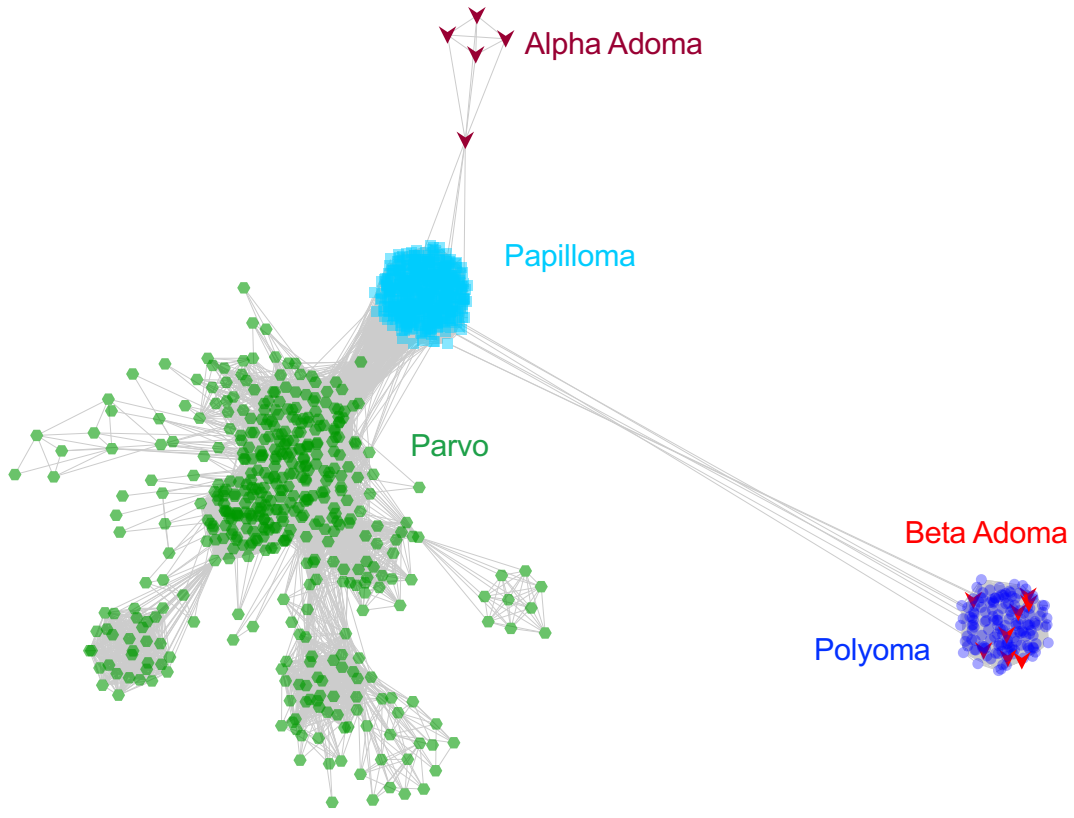

Figure 4: High stringency (E-10) sequence similarity network analysis of S3H proteins from indicated virus groups. An interactive Cytoscape file that incorporates virus names, accession numbers, and sequences is posted at https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoNetwork

It is thought that polyomaviruses and papillomaviruses tend to co-evolve with their animal hosts (*21, 22*). This is reflected in the fact that phylogeny within polyomavirus and papillomavirus genera recapitulates the phylogenetic relationships of host animals. Although adomavirus phylogenetic trees are currently sparse, the available data are consistent with adomavirus-host co-evolution similar to that of polyomaviruses and papillomaviruses (Supplemental Figure 4). For example, the phylogeny of adomavirus LO8 (adenain) proteins roughly correlates with the animal classes in which the sequences were found. In particular, alpha adomavirus sequences observed in RNA datasets for brook salamander and anole lizards are more closely related to one another than to fish-associated adomaviruses. The results suggest that alpha adomaviruses and beta adomaviruses have been evolving as independent lineages for at least half a billion years (Supplemental Figure 4).

Discovery of adintoviruses

Database searches for adomaviruses revealed a large number of contigs uniting an adenain homolog, penton- and hexon-like genes, and a retrovirus-like integrase gene (Figure 5, Supplemental Figure 5). We refer to this group of contigs as adintoviruses. A hallmark adintovirus feature is a large protein with a characteristic N-terminal RNA-binding zinc finger-like domain, a central papain/OTU1-like domain, an endonuclease VII-like domain, and a C-terminal type B DNA polymerase (PolB) domain. Network analysis of adintovirus PolB genes

divides them into two distinct clades (Figure 6). The predicted PolB proteins of a lineage we designate alpha show similarity to adenovirus PolB genes, while beta adintovirus PolB genes are distantly similar to bacteriophage PolB sequences. The appearance of two distinct clades is recapitulated in network analyses of predicted adintovirus hexon proteins (data not shown), suggesting the existence of two independent lineages.

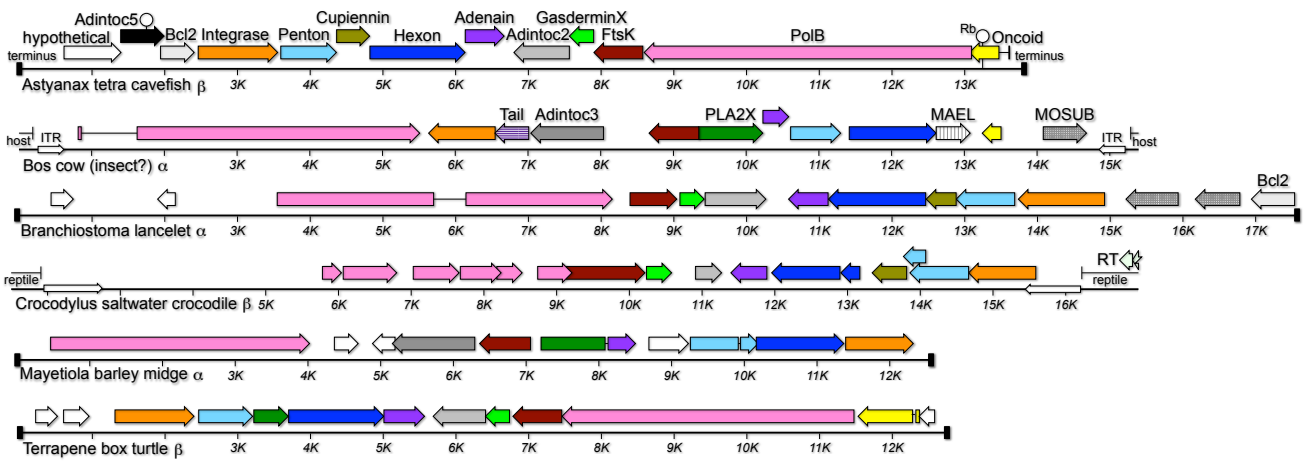

Figure 5: Picture maps of select adintovirus genomes. The figure uses the same color scheme as Figure 1. Supplemental Figure 5 shows additional examples of adintovirus genomes. Supplemental Table 1 lists accession numbers and host animal taxonomic designations.

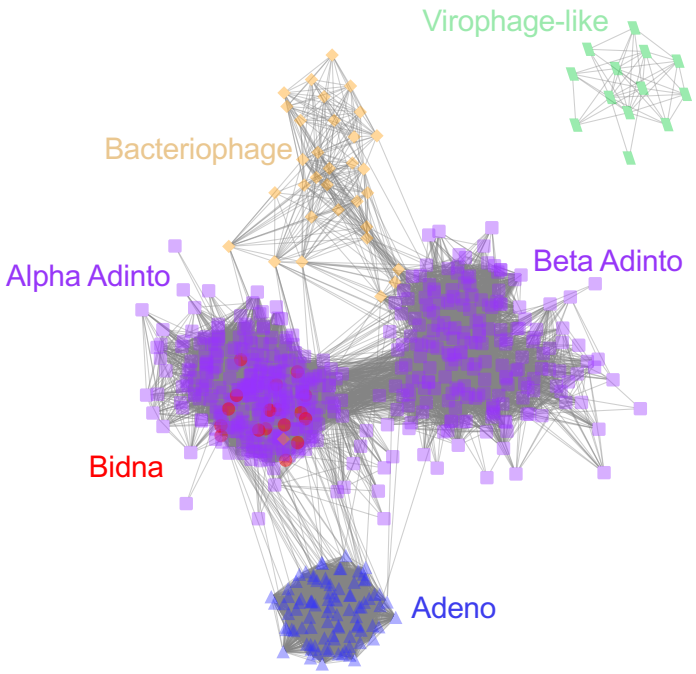

Figure 6: Moderate stringency (E-5) network analysis of PolB proteins. An interactive Cytoscape file that incorporates virus names, accession numbers, and sequences is posted at https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoNetwork

225 Apparently complete adintovirus genomes were re-assembled from high-throughput sequencing
226 datasets generated from a wide variety of animal types (Supplemental Table 1 and
227 https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoAll). In many cases, the consensus of
228 mapped reads ended abruptly, suggesting a linear, non-integrated viral genome. In some
229 instances, such as a barley midge (*Mayetiola destructor*) dataset, a complete linear adintovirus
230 genome could be assembled, but the dataset also contained a range of lower-coverage variant
231 sequences, some of which were flanked by inverted terminal repeats (ITRs) contiguous with
232 insect genomic DNA sequences. This observation suggests that the adintovirus integrase gene is
233 functional and mediates distinct integration events akin to those observed in virophages (*8*). A
234 subset of apparently integrated adintoviruses, such as contigs assembled from various crocodile
235 datasets, are heavily interspersed with frameshift and nonsense mutations (Figure 5). In many
236 cases, adintovirus-like sequences appear to be adjacent to or even infiltrated by endogenous
237 retroelements. These results indicate that some adintoviruses are endogenized into animal
238 germline DNA.

239 A majority of adintoviruses encode a homolog of the phospholipase A2 (PLA2) domain of
240 parvovirus VP1 virion proteins. The predicted adintovirus PLA2 homologs typically include a C-
241 terminal domain with similarity to adenovirus pX. Adintovirus "PLA2X" proteins also show
242 high probability HHpred hits for melittin and secreted PLA2 proteins found in bee and snake
243 venom. The fact that venom melittin and PLA2 act in concert (*23*) raises the interesting
244 hypothesis that some animal venom proteins might share common ancestry with membrane-
245 active viral capsid proteins.

246 Some adintoviruses encode homologs of the C-terminal regulatory domain of gasdermins, a
247 group of pore-forming proteins that serve as executioners in pyroptosis (a form of programmed
248 cell death) (*24*). Adintovirus homologs of a spider venom protein known as cupiennin were also
249 observed. Other predicted protein sequences were conserved among various adintoviruses but
250 did not show clear hits in BLAST or HHpred searches. We assigned these groups of unknown
251 adinto-conserved proteins numbered "Adintoc" names.

252 All classes of small DNA tumor viruses encode proteins harboring motifs, such as LXCXE, that
253 that are known to engage cellular retinoblastoma (Rb) and related tumor suppressor proteins (*2*).
254 Adenovirus E1A, papillomavirus E7, polyomavirus LT, and parvovirus NS3 oncogenes typically
255 encode the Rb-binding motif just upstream of a consensus casein kinase 2 acceptor motif
256 ((ST)XX(DE)). Some oncogenes, such as E1A, encode an additional conserved motif, referred to
257 as CR1, that resembles Rb pAB groove-interacting motifs ((DEN)(LIMV)XX(LM)(FY)) (*19,*
258 *25*). In general, these predicted Rb-interacting motifs are adjacent to potential zinc or iron-sulfur
259 binding motifs (typically, paired CXXC). Open reading frames encoding combinations of these
260 motifs are frequently observed in adintovirus contigs (Supplemental Figure 5). We refer to these
261 predicted proteins as "oncoids" due to their similarities to the known oncogenes of small DNA
262 tumor viruses. Predicted adintovirus homologs of anti-apoptotic proteins, such as Bcl2 and IAP,
263 were also observed. Surprisingly, oncoid-like genes were also observed in polintoviruses of
264 unicellular organisms (Supplemental Table 1, Supplemental Figure 6 and
265 https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoAll). These observations raise the
266 hypothesis that the oncogenes of modern small DNA tumor viruses descended from accessory
267 genes that first arose in viruses of early eukaryotes.

Adintoviruses were detected in diverse chordate datasets ranging from lancelets (*Branchiostoma belcheri*) to box turtles (*Terrapene mexicana triunguis*). A seemingly complete adintovirus genome was found in a transcriptomic dataset for a Mexican blind tetra cavefish (*Astyanax mexicanus*). Adintovirus transcripts were most abundant in cavefish head, kidney, and intestine samples and least abundant in muscle and whole embryo samples (Supplemental Table 4). In contrast to the lancelet and box turtle datasets, which showed a high degree of intra-host adintovirus sequence diversity at varying abundance, the cavefish adintovirus sequences were relatively uniform. The observation is consistent with possible bottlenecking of a single adintovirus strain into the sampled population of aquarium fish. Cavefish could provide a tractable small vertebrate model for adintovirus infection.

The only obvious examples of adintoviruses in mammal datasets were found in a PacBio-based WGS survey of bovine lung tissue. Sequences outside the ITRs were highly diverse but in a few instances showed BLASTN similarity to genomic DNA sequences of various beetles, including *Tribolium castaneum* (a flour beetle that commonly infests cattle feed). It is possible that adintovirus sequences found in some datasets represent environmental contaminants, as opposed to an infection of the organism that was the subject of the sequencing effort.

Identification of exotic parvoviruses

BLAST searches using alpha adintovirus PolB sequences consistently returned high-likelihood matches to the PolB genes of an emerging group of bipartite parvo-like viruses known as bidnaviruses (*26*). SRA searches for additional examples of bidna-class PolB genes revealed bidnavirus-like contigs in datasets for false wolf spider (*Tengella perfuga*) and Tasmanian devil (*Sarcophilus harrisii*) feces (Figure 7). The datasets were searched for additional contigs encoding the same inverted terminal repeat sequences (which are conserved between the two segments of known bidnaviruses) and for contigs encoding homologs of previously reported bidnavirus proteins. Second segments were not detected, suggesting that the *Tengella* and *Sarcophilus* bidnaviruses may be monopartite.

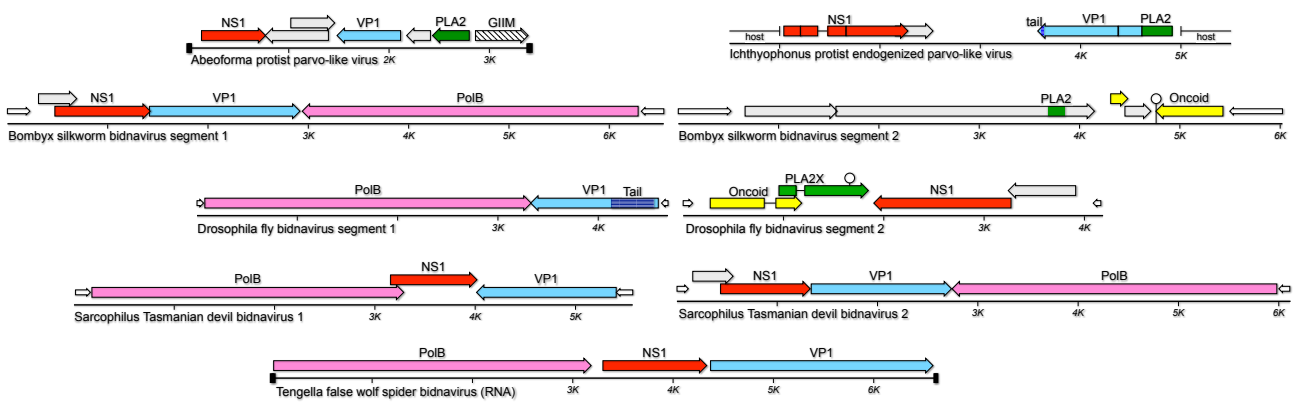

Figure 7. Parvovirus genome maps. Figure uses the same color scheme as Figure 1. GIIM indicates similarity to group II intron maturases. Accession numbers are listed in Supplemental Table 1.

295 Sequence queries in GenBank nr, WGS, and TSA databases revealed a large and highly diverse
296 range of parvovirus NS1-like genes in all major animal phyla. With the exception of parvovirus-
297 like contigs found in datasets for *Abeoforma whisleri* and *Ichthyophonus hoferi* (unicellular
298 eukaryotes that are thought to be closely related to animals), parvovirus NS1 sequences were not
299 detected in datasets for non-animal eukaryotes. Interestingly, the *Abeoforma* contig encodes a
300 protein with similarity (HHPred probability 82%) to the reverse transcriptase domain of group II
301 intron maturases.

302 **Analysis of S3H genes**

303 Polyomavirus LT, papillomavirus E1, and adomavirus EO1 replicase proteins have an origin-
304 binding domain that shows clear structural similarity to the HUH "nickase" endonuclease
305 domain of parvovirus NS1 proteins (*3, 17*). The replicase genes of CRESS viruses share the
306 same HUH-S3H domain architecture. Many CRESS viruses encode an ORF "overprinted" in the
307 +1 frame of the Rep ORF. Similarly, overprinted ORFs are often observed in parvovirus NS1
308 genes as well as in polyomavirus LT genes (*27*).

309 We did not detect examples of adintovirus-specific S3H genes. Although many virophage-class
310 viruses encode proteins with S3H domains (*28*), our searches confirm that the S3H proteins of
311 virophage-like viruses lack detectable HUH endonuclease-like domains and lack overprinted
312 ORFs. Furthermore, virophage-class S3H proteins did not cluster with small DNA tumor virus
313 S3H proteins in network analyses. In contrast, similarity was observed between a small subset of
314 CRESS virus Rep proteins and the replicase proteins of parvoviruses and papillomaviruses
315 (Figure 8). These observations support the previously proposed view that CRESS, polyoma,
316 papilloma, and parvovirus replicase genes share more recent common ancestry with one another
317 than with the S3H genes of other known virus lineages (*3*).

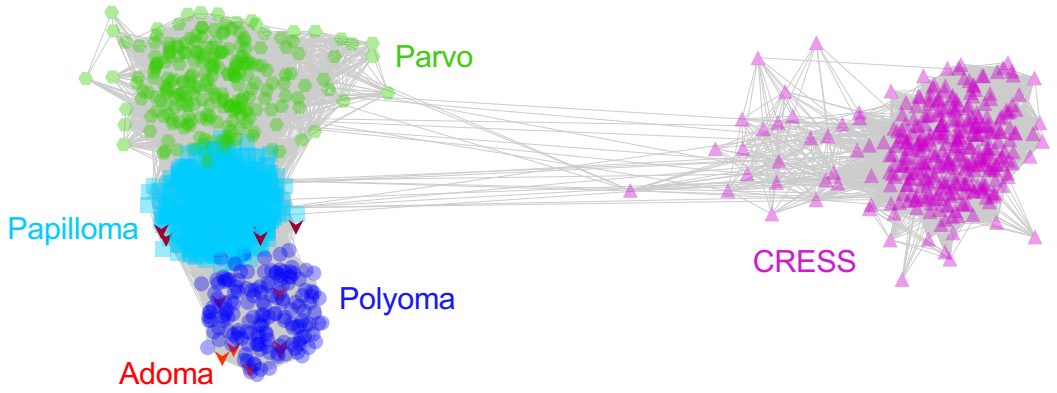

Figure 8: Network analysis of S3H replicase proteins. An EFI-EST analysis with a cutoff of E-5 was performed on small DNA tumor virus and select CRESS virus replicase proteins. A Cytoscape file indicating virus names and accession numbers is posted at https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoNetwork

318

319 **Discussion**

320 In this report, we began by identifying the hexon and penton genes of the emerging adomavirus
321 family. We demonstrated that the adomavirus LO4-8 array is homologous to the adenovirus late

322 gene array that encodes most of the virion structural proteins. The ability of adomavirus virion
323 proteins to assemble into irregular "virus-like" particles should facilitate the development of
324 recombinant vaccines against adomaviruses.

325 We identified two distinct lineages of adomaviruses that appear to have been independently co-
326 evolving with their vertebrate hosts for at least half a billion years (Supplemental Figure 4).
327 Within both of these adomavirus lineages, there are members associated with commercially
328 important fish species. Surprisingly three closely related alpha adomavirus-like segments were
329 observed in RNA sequencing datasets for terrestrial vertebrates. It will be important to apply
330 emerging higher-throughput search algorithms, such as Mash Screen (*29*), to exhaustively search
331 for adomaviruses in deep sequencing datasets for other terrestrial vertebrates, particularly
332 humans.

333 Our searches for adomaviruses serendipitously revealed an apparently older lineage of animal-
334 tropic viruses that we name adintoviruses. Adintoviruses encode a number of genes that appear
335 to be homologs of membrane-active proteins found in animal venom. These include bee and
336 snake venom PLA2 and melittin, as well as a spider venom protein called cupiennin. In some
337 cases, these proteins also score as weakly similar to bacterial colicin proteins or bacteriophage
338 holin proteins, which play a role in disrupting bacterial membranes during the infectious entry
339 process. We speculate that some types of animal venom genes may be endogenized membrane
340 active virion protein genes. It also appears that alpha adintoviruses share relatively recent
341 common ancestry with some segments of bracoviruses (Supplemental Figure 5), a group of
342 endogenized virus-like elements that parasitoid wasps inject alongside their eggs (*30*).

343 It has generally been assumed that the functionally similar oncogenes found in papillomaviruses,
344 polyomaviruses, parvoviruses, and adenoviruses arose through convergent evolution or through
345 horizontal gene transfer between virus families (*2*). Although small DNA tumor virus oncogenes
346 show extremely low overall sequence similarity, they can be roughly defined based on the
347 presence of short linear motifs. Many adintoviruses encode candidate "oncoid" proteins with
348 these motifs. Unexpectedly, it appears that some virophage-class viruses of unicellular organisms
349 also encode oncogene-like proteins (Supplemental Figure 6). These observations raise the
350 hypothesis that the oncogenes of modern small DNA tumor viruses might have descended from a
351 class of accessory genes that evolved before the appearance of multicellular animals.

352 Although the current study indicates clear homologies between adenovirus, adomavirus, and
353 adintovirus virion proteins, the origins of polyomavirus and papillomavirus virion proteins
354 remain unclear. It has previously been suggested that the major capsid proteins the smaller tumor
355 virus families might have arisen from the capsid proteins of CRESS and/or ssRNA viruses (*3*).
356 This would have involved a complex shift from the simple T=1 icosahedral geometry of CRESS
357 virions, in which the strands of the core ß-jellyroll of the major capsid protein are parallel to the
358 lumen of the virion, to the more complex T=7 perpendicular jellyroll organization seen in
359 papillomaviruses and polyomaviruses (*1*). Furthermore, CRESS viruses typically have genomes
360 less than 3 kb in size, while papillomaviruses and polyomaviruses are typically 5-8 kb in size.
361 Although a few CRESS viruses with 5-9 kb genomes have recently been reported, the larger
362 CRESS virus genomes each encode multiple parallel-jellyroll capsid protein paralogs or
363 unrelated non-jellyroll filamentous virion proteins (*31*). Thus, there does not appear to be a clear
364 precedent for the proposed parallel-to-perpendicular reorientation of a single-jellyroll subunit
365 among known CRESS viruses. We propose a simpler model in which the pentameric

366 papillomavirus and polyomavirus capsid proteins were derived from the perpendicular single-
367 jellyroll penton proteins of adintovirus-like ancestors. The observation that the inferred penton
368 protein of marbled eel adomavirus can assemble into roughly spherical ~50 nm structures
369 carrying non-specifically encapsidated dsDNA raises the possibility that minimal evolutionary
370 adaptations would be required to construct a 72-penton structure capable of housing a
371 polyomavirus or papillomavirus genome. This model could be extended to the splayed-jellyroll
372 capsomers of parvoviruses. One line of support for this idea is the observation that parvovirus
373 virion proteins typically include an N-terminal domain closely resembling adintovirus PLA2X
374 proteins. Similarly, a small subset of parvoviruses appear to encode capsid protein domains with
375 distant similarity to predicted adintovirus tail-like proteins (Figure 7) (*32*).

376 To further explore the hypothesis that papillomavirus, polyomavirus, and parvovirus virion
377 proteins were derived from adintovirus-like ancestors, extremely low stringency network
378 analyses were performed for various classes of major and minor capsid proteins. The analysis
379 rests on the assumption that a small subset of members of each virus family might better reflect
380 the primary sequence of a common ancestor, which would be revealed by linkages between a
381 few members of different clusters. Aside from the previously described distant similarity
382 between the parvovirus PLA2 domain and the Cap genes of small subset of non-animal CRESS
383 viruses (*33, 34*), no clear connections were observed between groups of oma-class virion
384 proteins and CRESS virus capsid proteins. The low-stringency analysis does, however, suggest
385 possible similarities between papillomavirus, polyomavirus, and parvovirus major capsid
386 proteins and the penton proteins of adenoviruses, adomaviruses, and adintoviruses (Supplemental
387 Figure 7). Possible distant similarity was also observed for some combinations of minor capsid
388 proteins (Supplemental Figure 8).

389 A hypothetical framework that could account for the apparent similarities between extant small
390 DNA tumor viruses is presented in Figure 9. A testable prediction of this model is that
391 adintovirus PLA2X and adomavirus LO6, like the papillomavirus and polyomavirus minor
392 capsid proteins, will be found to directly interact with the luminal surface of pentons in the
393 assembled virion. We also predict that the fold of polyomavirus VP2 and papillomavirus L2
394 proteins may be found to resemble the folds of adintovirus PLA2X, GasderminX, or cupiennin
395 proteins. The model further predicts that adenovirus pX, in addition to its proposed role in
396 condensing adenovirus genomes (*10*), participates in destabilization of host cell membranes
397 during the infectious entry process in a manner similar to that of papillomavirus L2 proteins (*35,*
398 *36*).

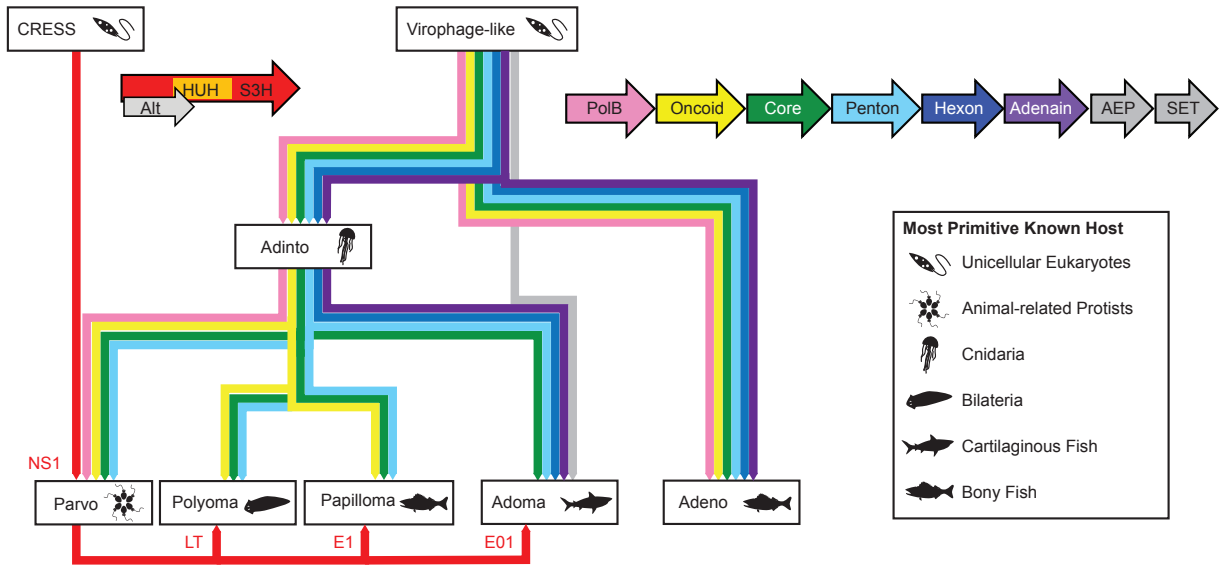

Figure 9: A hypothetical framework for small DNA tumor virus evolution.

At the outset of this study, adintoviruses were not recognizable as viruses in basic BLAST searches and time-consuming structure-guided approaches were required to reveal distant similarities to known proteins. An important implication is that there may be additional unrecognized families of animal viruses hiding in plain sight in sequence databases. To address this problem, it will be important to develop higher throughput methods for more sensitive structure-guided searches and for understanding which combinations predicted homologs co-occupy single contigs. It will be especially important to apply high throughput annotation pipelines, such as Cenote-Taker (*31*), that generate human-readable GenBank compliant maps. Deposition of these annotated new viral genomes in publicly accessible databases will be critical for expanding our understanding of the animal virome.

Materials and Methods

Sample Collection and cell culture

A red discus cichlid (*Symphysodon discus*) was purchased at a pet shop in Gainesville, Florida. The fish was moribund and showed erythematous skin lesions. Propagation of the discus adomavirus in cell culture was attempted by overlaying skin tissue homogenates on Grunt Fin (GF) and Epithelioma Papulosum Cyprini cell lines (ATCC). Neither cytopathic effects nor qPCR-based detection of viral replication were observed during two blind passages of 14 days each.

Dr. Chiu-Ming Wen generously provided EK-1 cells (a Japanese eel kidney line) infected with the Taiwanese marbled eel adomavirus (*6*). The virus was propagated by inoculation of supernatants from the infected culture into uninfected EK-1 cells cultured at room temperature in

422 DMEM with 10 % fetal calf serum. Human embryonic kidney-derived 293TT cells were cultured
423 as previously described (*14*).

424 **Viral genome sequencing**

425 For the discus adomavirus, total DNA was extracted from a skin lesion and subjected to deep
426 sequencing. Marbled eel adomavirus virions were purified from lysates of infected EK-1 cells
427 using Optiprep gradient ultracentrifugation (*37*). DNA extracted from Optiprep gradient fractions
428 was subjected to rolling circle amplification (RCA, TempliPhi, GE Health Sciences). The
429 marbled eel adomavirus RCA products and discus total DNA were prepared with a Nextera XT
430 DNA Sample Prep kit and sequenced using the MiSeq (Illumina) sequencing system with 2 ×
431 250 bp paired-end sequencing reagents. In addition, the marbled eel adomavirus RCA product
432 was digested with *AclI* and *EcoRI* restriction enzymes and the resulting early and late halves of
433 the viral genome were cloned separately into the *AclI* and *EcoRI* restriction sites of pAsylum+.
434 The sequence of the cloned genome was verified by a combination of MiSeq and Sanger
435 sequencing. The clones will be made available upon request.

436 For the arowana adomavirus, overlapping PCR primers were designed based on WGS accession
437 numbers LGSE01029406, LGSE01031009, LGSE01028643, LGSE01028176, and
438 LGSE01030049 (*16*). PCR products were subjected to primer walking Sanger sequencing.

439 **Marbled eel adomavirus transcript analysis, late ORF expression, and virion purification**

440 RNAseq reads reported by Wen et al (*6*) were aligned to the marbled eel adomavirus genome
441 using HISAT2 version 2.0.5 (*38*) with the following options: "--rna-strandness FR --dta --no-
442 mixed --no-discordant". Integrated Genome Viewer (IGV) version 2.4.9 (*39*) was used to
443 determine splice junctions and their depth of coverage. Additional validation was performed by
444 visual inspection using CLC Genomics Workbench 12.

445 Codon-modified expression constructs encoding the marbled eel adomavirus LO1-LO8 proteins
446 were designed according to a modified version of a previously reported algorithm
447 (https://github.com/BUCK-LCO-NCI/Codmod_as_different_as_possible)(*40*). 293TT cells were
448 transfected with LO expression constructs for roughly 48 hours. Cells were lysed in a small
449 volume of PBS with 0.5% Triton X-100 or Brij-58 and Benzonase Dnase/Rnase (Sigma)(*41*).
450 After one hour of maturation at neutral pH, the lysate was clarified at 5000 x g for 10 min. The
451 clarified lysate was loaded onto a 15-27-33-39-46% Optiprep gradient in PBS with 0.8 M NaCl.
452 Gradient fractions were collected by bottom puncture of the tube and screened by PicoGreen
453 dsDNA stain (Invitrogen), BCA, or SDS-PAGE analysis. Electron microscopic analysis was
454 performed by spotting several microliters of Optiprep fraction material (or, in some instances,
455 particles exchanged out of Optiprep using agarose gel filtration) onto carbon support film on
456 Cu/Ni/Au grids, followed by staining with 0.5% uranyl acetate.
457

458 **Discovery of viral sequences in NCBI databases**

459 Compilations of sequences use in this study are posted at the following link:
460 https://ccrod.cancer.gov/confluence/display/LCOTF/Adinto. Papillomavirus E1 sequences were
461 downloaded from PaVE https://pave.niaid.nih.gov (*42*). Polyomavirus LT sequences were
462 downloaded from PyVE https://ccrod.cancer.gov/confluence/display/LCOTF/Polyomavirus (*22*).
463 Parvovirus NS1 proteins and S3H proteins of virophage-like viruses were compiled from

464 multiple databases, including RefSeq, WGS, and TSA, using TBLASTN searches. Adenovirus,
465 virophage, and bacteriophage PolB sequences were downloaded from GenBank nr using
466 DELTA-BLAST searches (*43*) with alpha or beta adintovirus PolB proteins as bait.

467 SRA datasets for fish, amphibians, and reptiles were searched using DIAMOND (*44*) or NCBI
468 SRA Toolkit (http://www.ncbi.nlm.nih.gov/books/NBK158900/) in TBLASTN mode using
469 adomavirus protein sequences as the subject database or query, respectively. Reads with
470 similarity to the baits were collected and subjected to BLASTX searches against a custom library
471 of viral proteins representing adomaviruses and other small DNA tumor viruses. SRA datasets of
472 interest were subjected to de novo assembly using the SPAdes suite (*45, 46*) or Megahit (*47, 48*).
473 Contigs encoding virus-like proteins were identified by TBLASTN searches against adomavirus
474 protein sequences using Bowtie (*49*). The candidate contigs were validated using the CLC
475 Genomics Workbench 12 align to reference function.

476 Adintoviruses were extracted from WGS and TSA datasets using TBLASTN searches with S3H
477 genes of various small DNA tumor viruses or the PolB, adenain, or PLA2 genes of Nephila orb-
478 weaver spider alpha adintovirus GFKT014647032. Predicted protein sequences were
479 automatically extracted from contigs of interest using getorf (http://bioinfo.nhri.org.tw/cgi-
480 bin/emboss/getorf) (*50*). Proteins were clustered using EFI-EST (https://efi.igb.illinois.edu/efi-
481 est/)(*51, 52*) and displayed using Cytoscape v3.7.1 (*53*). Multiple sequence alignments were
482 constructed using MAFFT (https://toolkit.tuebingen.mpg.de/#/tools/mafft). Individual or aligned
483 protein sequences were subjected to HHpred searches
484 (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) (*54*) of PDB, Pfam-A, NCBI Conserved
485 Domains, and PRK databases.

486 Contigs were annotated using Cenote-Taker (*31*) with an iteratively refined library of conserved
487 adintovirus protein sequences. Maps were drawn using MacVector 17 software.

488 The phylogenetic analysis shown in Supplemental Figure 4 was performed using Phylogeny.fr
489 (*55*). Host animal divergence times were obtained from TimeTree of Life
490 http://www.timetree.org (*56*).

491 **Mass Spectrometry**

492 Optiprep-purified marbled eel adomavirus virions were precipitated with trichloroacetic acid. A
493 1 ml sample was treated with 100 µl of 0.15% deoxycholic acid and incubated at room
494 temperature for 10 minutes. 100 µl of 100% TCA was then added and the sample was vortexed
495 and incubated on ice for 30 minutes. Following the incubation, the sample was centrifuged at
496 10,000 x g for 10 minutes at 4°C. The supernatant was removed, and the remaining pellet was
497 washed with ice-cold acetone to remove residual TCA. The protein pellet was solublized with
498 NuPAGE Sample Buffer + 5% BME (Sigma) and run on a 10-12% Bis-Tris MOPS gel
499 (Thermo). The protein bands were visualized using InstantBlue (Expedeon). Thirteen gel bands
500 were individually excised and placed into 1.5 ml Eppendorf tubes. The gel bands were sent to the
501 National Cancer Institute in Fredrick, Maryland where they were de-stained, digested with
502 trypsin, and processed on a Thermo Fisher Q Exactive HF Mass Spectrometer. Thermo Proteome
503 Discoverer 2.2 software was used for initial protein identification. The uninterpreted mass
504 spectral data were also searched against *Anguilla* proteins (Swiss-Prot and TrEMBL database

containing 105,268 proteins), Bos taurus proteins (Swiss-Prot and TrEMBL database containing 48,288 proteins), a common contaminants database (cRAPome), and translated marbled eel adomavirus ORFs. Further analysis was conducted using Protein Metrics Biopharma software to identify modifications missed in initial analyses.

Ethics Statement

All animal tissue samples were received as diagnostic specimens collected for pathogen testing and disease investigation purposes.

Data Availability

Sequences generated from this study are available via the accession numbers listed in Supplemental Table 1. Additional annotated sequence files are posted at https://ccrod.cancer.gov/confluence/display/LCOTF/Adinto

Acknowledgments

The authors are indebted to Eugene Koonin and Natalya Yutin for their generous guidance and for the spirited discussions that inspired us to pursue this study. We are particularly grateful to them for sharing their observation that adomaviruses encode a recognizable adenain homolog and their discovery of adomavirus sequences in the arowana WGS datasets. The authors are also grateful to Lisa Jenkins for her extensive technical guidance on analyzing mass spectrometric data. We thank Karl Münger for useful discussions, including advice about oncogene sequence motifs.

524 **References**

525

526

527 1. M. Krupovic, E. V. Koonin, Multiple origins of viral capsid proteins from cellular
528 ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410 (2017).
529 2. R. F. de Souza, L. M. Iyer, L. Aravind, Diversity and evolution of chromatin proteins
530 encoded by DNA viruses. *Biochim Biophys Acta* **1799**, 302-318 (2010).
531 3. E. V. Koonin, V. V. Dolja, M. Krupovic, Origins and evolution of viruses of eukaryotes:
532 The ultimate modularity. *Virology* **479-480**, 2-25 (2015).
533 4. T. Mizutani *et al.*, Novel DNA virus isolated from samples showing endothelial cell
534 necrosis in the Japanese eel, Anguilla japonica. *Virology* **412**, 179-187 (2011).
535 5. S. Okazaki *et al.*, Detection of Japanese eel endothelial cells-infecting virus in Anguilla
536 japonica elvers. *J Vet Med Sci* **78**, 705-707 (2016).
537 6. C. M. Wen, M. M. Chen, C. S. Wang, P. C. Liu, F. H. Nan, Isolation of a novel
538 polyomavirus, related to Japanese eel endothelial cell-infecting virus, from marbled eels,
539 Anguilla marmorata (Quoy & Gaimard). *J Fish Dis*, (2015).
540 7. J. A. Dill, A. C. Camus, J. H. Leary, T. F. F. Ng, Microscopic and Molecular Evidence of
541 the First Elasmobranch Adomavirus, the Cause of Skin Disease in a Giant Guitarfish,
542 Rhynchobatus djiddensis. *MBio* **9**, (2018).
543 8. M. G. Fischer, T. Hackl, Host genome integration and giant virus-induced reactivation of
544 the virophage mavirus. *Nature* **540**, 288-291 (2016).
545 9. J. Logan, T. Shenk, Adenovirus tripartite leader sequence enhances translation of
546 mRNAs late after infection. *Proc Natl Acad Sci U S A* **81**, 3655-3659 (1984).
547 10. G. R. Nemerow, P. L. Stewart, V. S. Reddy, Structure of human adenovirus. *Curr Opin*
548 *Virol* **2**, 115-121 (2012).
549 11. C. L. Moyer, E. S. Besser, G. R. Nemerow, A Single Maturation Cleavage Site in
550 Adenovirus Impacts Cell Entry and Capsid Assembly. *J Virol* **90**, 521-532 (2015).
551 12. A. Ruzindana-Umunyana, L. Imbeault, J. M. Weber, Substrate specificity of adenovirus
552 protease. *Virus Res* **89**, 41-52 (2002).
553 13. C. Vragniau *et al.*, Studies on the Interaction of Tumor-Derived HD5 Alpha Defensins
554 with Adenoviruses and Implications for Oncolytic Adenovirus Therapy. *J Virol* **91**,
555 (2017).
556 14. C. B. Buck, D. V. Pastrana, D. R. Lowy, J. T. Schiller, Efficient intracellular assembly of
557 papillomaviral vectors. *J Virol* **78**, 751-757 (2004).
558 15. C. B. Buck, C. D. Thompson, Y. Y. Pang, D. R. Lowy, J. T. Schiller, Maturation of
559 papillomavirus capsids. *J Virol* **79**, 2839-2846 (2005).
560 16. C. Bian *et al.*, The Asian arowana (Scleropages formosus) genome provides new insights
561 into the evolution of an early lineage of teleosts. *Sci Rep* **6**, 24501 (2016).
562 17. A. B. Hickman, D. R. Ronning, R. M. Kotin, F. Dyda, Structural unity among viral origin
563 binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep.
564 *Mol Cell* **10**, 327-337 (2002).
565 18. L. M. Iyer, E. V. Koonin, D. D. Leipe, L. Aravind, Origin and evolution of the archaeo-
566 eukaryotic primase superfamily and related palm-domain proteins: structural insights and
567 new members. *Nucleic Acids Res* **33**, 3875-3896 (2005).

568 19. M. Gouw *et al.*, The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res*
569 **46**, D428-D434 (2018).
570 20. P. An, M. T. Saenz Robles, J. M. Pipas, Large T antigens of polyomaviruses: amazing
571 molecular machines. *Annu Rev Microbiol* **66**, 213-236 (2012).
572 21. A. Rector *et al.*, Ancient papillomavirus-host co-speciation in Felidae. *Genome biology* **8**,
573 R57 (2007).
574 22. C. B. Buck *et al.*, The Ancient Evolutionary History of Polyomaviruses. *PLoS pathogens*
575 **12**, e1005574 (2016).
576 23. W. Vogt, P. Patzer, L. Lege, H. D. Oldigs, G. Wille, Synergism between phospholipase A
577 and various peptides and SH-reagents in causing haemolysis. *Naunyn Schmiedebergs*
578 *Arch Pharmakol* **265**, 442-454 (1970).
579 24. H. Dubois *et al.*, Nlrp3 inflammasome activation and Gasdermin D-driven pyroptosis are
580 immunopathogenic upon gastrointestinal norovirus infection. *PLoS pathogens* **15**,
581 e1007709 (2019).
582 25. J. M. Pipas, Common and unique features of T antigens encoded by the polyomavirus
583 group. *J Virol* **66**, 3979-3985 (1992).
584 26. M. Krupovic, E. V. Koonin, Evolution of eukaryotic single-stranded DNA viruses of the
585 Bidnaviridae family from genes of four other groups of widely different viruses. *Sci Rep*
586 **4**, 5347 (2014).
587 27. J. J. Carter *et al.*, Identification of an overprinting gene in Merkel cell polyomavirus
588 provides evolutionary insight into the birth of viral genes. *Proc Natl Acad Sci U S A* **110**,
589 12744-12749 (2013).
590 28. N. Yutin, S. Shevchenko, V. Kapitonov, M. Krupovic, E. V. Koonin, A novel group of
591 diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol* **13**, 95
592 (2015).
593 29. B. D. Ondov *et al.*, Mash Screen: High-throughput sequence containment estimation for
594 genome discovery. *bioRxiv*, 557314 (2019).
595 30. M. R. Strand, G. R. Burke, Polydnaviruses: Nature's Genetic Engineers. *Annu Rev Virol*
596 **1**, 333-354 (2014).
597 31. M. J. Tisza *et al.*, Discovery of several thousand highly diverse circular DNA viruses.
598 555375 (2019).
599 32. J. J. Penzes *et al.*, Molecular characterization of a lizard adenovirus reveals the first
600 atadenovirus with two fiber genes and the first adenovirus with either one short or three
601 long fibers per penton. *J Virol* **88**, 11304-11314 (2014).
602 33. B. Xu *et al.*, Hybrid DNA virus in Chinese patients with seronegative hepatitis
603 discovered by deep sequencing. *Proc Natl Acad Sci U S A* **110**, 10264-10269 (2013).
604 34. S. N. Naccache *et al.*, The perils of pathogen discovery: origin of a novel parvovirus-like
605 hybrid genome traced to nucleic acid extraction spin columns. *J Virol* **87**, 11966-11977
606 (2013).
607 35. I. Aydin *et al.*, A central region in the minor capsid protein of papillomaviruses facilitates
608 viral genome tethering and membrane penetration for mitotic nuclear entry. *PLoS*
609 *pathogens* **13**, e1006308 (2017).
610 36. S. K. Campos, Subcellular Trafficking of the Papillomavirus Genome during Initial
611 Infection: The Remarkable Abilities of Minor Capsid Protein L2. *Viruses* **9**, (2017).
612 37. A. Peretti, P. C. FitzGerald, V. Bliskovsky, C. B. Buck, D. V. Pastrana, Hamburger
613 polyomaviruses. *J Gen Virol* **96**, 833-839 (2015).

38. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360 (2015).
39. J. T. Robinson, H. Thorvaldsdottir, A. M. Wenger, A. Zehir, J. P. Mesirov, Variant Review with the Integrative Genomics Viewer. *Cancer Res* **77**, e31-e34 (2017).
40. D. V. Pastrana *et al.*, Reactivity of human sera in a sensitive, high-throughput pseudovirus-based papillomavirus neutralization assay for HPV16 and HPV18. *Virology* **321**, 205-216 (2004).
41. C. B. Buck, C. D. Thompson, Production of papillomavirus-based gene transfer vectors. *Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al* **Chapter 26**, Unit 26.21 (2007).
42. K. Van Doorslaer *et al.*, The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res* **45**, D499-d506 (2017).
43. G. M. Boratyn *et al.*, Domain enhanced lookup time accelerated BLAST. *Biology direct* **7**, 12 (2012).
44. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60 (2015).
45. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455-477 (2012).
46. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome research* **27**, 824-834 (2017).
47. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
48. D. Li *et al.*, MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11 (2016).
49. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
50. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).
51. J. A. Gerlt *et al.*, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta* **1854**, 1019-1037 (2015).
52. R. Zallot, N. O. Oberg, J. A. Gerlt, 'Democratized' genomic enzymology web tools for functional assignment. *Curr Opin Chem Biol* **47**, 77-85 (2018).
53. P. Shannon *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).
54. L. Zimmermann *et al.*, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **S0022-2836**, 30587-30589 (2017).
55. A. Dereeper *et al.*, Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**, W465-469 (2008).
56. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution* **34**, 1812-1819 (2017).